Bradford Hill criteria with totally useful and not at all silly examples

Jesse Brunner

2022-06-02

Bradford Hill's criteria are, I think, terribly useful in thinking about the evidence for causal relationships. They were borne from, and are still commonly applied in, an epidemiological setting, but we can use them more broadly. It is worth noting that the many people who quibble with BHC take issue with that last word, "criteria." It is not as if one must 'tick' all of the boxes to demonstrate causality, nor is it certain that if one did they would be right[1]. Still, they are quite useful if, like all tools[2], they are used for the right sorts of problems and one is mindful about where the sharp bits are.

All that said, students are often a bit mystified by these criteria, conflating some, splitting hairs elsewhere, and generally feeling pretty muddled about it all. So let me offer some examples. And let me encourage you to come up with your own!

[1] Perhaps it should be Bradford Hill's *characteristics*?
[2] Bradford Hill's *tools*?

Plausibility

> *A plausible mechanism between cause and effect is helpful*[3]

Plausible means reasonable or probable. In this case of BHC we're talking about the plausibility of the mechanism(s) invoked by an assertion. For instance, consider which of these two scenarios is more plausible.

1. Parents put money under their kids' pillows and take the teeth left there, maintaining a weird, but sweet tradition.
2. An invisible fairy somehow knows when and where teeth have been lost, is quite keen on collecting them[4], and is willing to pay good money[5] for them, even though s/he can apparently sneak in and out of a kid's bedroom without being noticed.

[3] This and others lifted straight from Wikipedia. Can I do that? I just did. But don't *you* do that in your papers!

[4] For what nefarious reasons we do not know!
[5] Seriously!

Strength of association or effect size

> *A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.*

Does the putative cause lead to a small or large change in the effect? This has nothing to do with sample size—that is more relevant to our confidence in the robustness of the findings—but rather with the magnitude of change (e.g., in an experiment) or difference (e.g., between groups, populations) or statistical association (e.g., in an epidemiological survey). For instance, my wearing my lucky socks might indeed have improved my organic chemistry test score on the three occasions I tried it, but the difference, a mere three percentage points on average

above my normal scores, suggests a small effect size that could easily be explained by other things[6].

[6] I *did* carbo-load the nights before!

Consistency (reproducibility) and Coherence

> *Consistent findings observed by different persons in different places with different samples strengthens the likelihood of [a causal] effect.*
>
> *Coherence between epidemiological and laboratory findings increases the likelihood of [a causal] effect.*

Consistency and coherence are often confused. Consistency is about the reproducibility of a pattern whereas coherence is about the unity of different sorts of studies, approaches, or measures. For instance, we love stories of college dropouts starting tech startups that become incredibly successful, and some have even said dropping out was instrumental in their success (e.g., not having a "plan B" was motivating), but the statistics show that the vast majority of people that do not finish their degree end up in lower-paying jobs suggests that the dropout-to-riches effect is inconsistent at best. Moreover, studies have shown that dropouts have smaller professional networks, higher debt-to-earning ratios, and even higher levels of circulating stress hormones, all of which paints a coherent picture of how dropping out makes the road to success even rockier and less likely[7].

[7] OK, I made this all up, but still, stay in school!

Specificity

> *Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.*

Specificity is about the uniqueness or distinctness of the group(s) or places or times for which the relationship applies (and does not apply). For instance, the fact that Tauruses are no more likely to experience a "financial windfall" on those days when our horoscopes predict it than any other sign of the zodiac suggests the effects of the stars are at best pretty non-specific. However, those people that consider themselves "lucky" *are* more likely to hear about new job opportunities, be promoted, and have a larger professional network than those who consider themselves "unlucky,"[8] which suggests there may be some causal relationship between perceptions of luckiness, or at least the appreciation of luck, and financial success[9].

[8] I.e., the effect is specific to "lucky" people.

[9] I'm making up the details, but there has indeed been a lot of science on what makes people "lucky".

Temporality

> *The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).*

Temporality can feel a bit trivial, but when the cause does *not* come before the effect, that can be a very strong signal that maybe the relationship isn't causal. I

always think about the saying, "when the dog runs toward you, say 'come!' " Or perhaps this is clearer: The Teletubbies have been blamed for many problems in society, but the fact that rates of autism began to increase in the early 1990s before the Teletubbies first aired in 1997, suggests the relationship is not causal. (Nope. I'm not making up this causal assertion!)

That said, just because the cause came before the effect does not mean it makes temporal sense. The timing has to make sense given the system, mechanism, and so on[10]. Sure, my kids, when they were babies, *did* go to sleep after I bounced them on the exercise ball while singing "Three little birds" to them, but sometimes it was right away and sometimes it was a *long* time after. I'm not sure if my singing and bouncing really caused them to sleep[11].

[10] This is one of the problems with fortune cookie predictions like "Your luck will improve." OK, but when?! Within the hour? Next week? Since there is no real mechanism specified it's hard to know what we should expect if there was a causal relationship.

[11] Getting babies to sleep makes everyone a bit superstitious. Some people get lucky with a hot streak and then write a book telling everyone else they should do the same thing they did. I should have written a book on my "Bouncy ball method" and made millions from other sleep-deprived parents!

Analogy

> *The use of analogies or similarities between the observed association and any other associations.*

Analogy comes from comparing similar things. When thinking about evidence, comparing comparable things can be useful, especially for showing that a mechanism that works in one place might be relevant in another. That is, what we learn in one system[12] (or place or time) may have lessons relevant to another. For instance, while we do not know whether the QAnon movement will remain cohesive and unified over time, the continued schisms in the analogous Flat Earth movement and prior UFOlogist groups suggests that as it grows it will inevitably splinter. Like literary analogies, however, you should always be mindful of extending it too far.

Note that finding similarities in terminology and wording does not necessarily indicate a similarity in the things themselves. Consider the numerous spiritual analogies made to quantum physics (e.g., matter is just energy, resonant frequencies, etc.) by spiritualists who do not actually understand quantum physics, to the deep annoyance of actual physicists!

[12] Often, we use experiments with mice and rats to learn by analogy what *might* be causal in humans. We're all mammals, after all, although results often fail to translate for a variety of reasons.

Biological gradient

> _ Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence._

This one is my favorite[13]: the dose-response relationship! We usually expect that if a little bit of exposure causes a little bit of response, then more exposure should cause more response. For instance, regions where more money is spent on advertising pet products tend to have larger numbers of overweight pets[14]. This is *probably* because, like me, our pets get hungry when they see food advertisements on TV. Either way is supports the assertion that advertising causes pet obesity.

There are caveats, however. Sometimes there are ranges of exposures where things change pretty rapidly, as if there were a threshold. Also, if you're already

[13] It comes with its own theme song!

[14] Yes, I just made this up. But it's probably true!

in the region of exposures where responses always or never happen, adding a bit more (or less) exposure might not do very much. Alternatively, you might end up in a different regime[15]. Feed your pet goldfish far too much and you'll notice it stops eating[16].

[15] Science-speak for a set of conditions where certain processes occur.

[16] And swimming!

Experiment

Occasionally it is possible to appeal to experimental evidence[17].

[17] That one is from BH himself in the original paper.

We are mostly familiar with experiments, but it is worth noting that an experiment involves an *intentional* manipulation. It is not enough just to observe groups of people or whatever that already differ in some interesting way, the experimenter must somehow intervene. Often this is adding or removing some factor of interest (e.g., tying one hand behind subjects backs to see if people that use their left hands really are evil; "sinister" comes from the Latin for "on the left side"). Ideally the treatment is applied randomly, so as to avoid bias (e.g., not just applied to those evil lefties) and has some controls (e.g., tying the dominant and non-dominant hand behind the back as well as having some subjects with sham knots to check for the effect of the rope around one's wrist). Experimental designs can be simple or complex, but they always involve a manipulation[18].

Note that sometimes the world can intervene in a manner akin to an experiment. We call these "natural experiments." They can be analyzed similarly to intentional experiments, but it is important to note that biases could creep in as to how or where or to whom the manipulation was applied. For instance, historical culling of left-handed people from populations might seem a terrific natural experiment to see if less evil occurred soon after, but it would be important to note that everyone might have been on their best behavior after that particular intervention, no matter their dominant hand!

[18] The advantage of experiments is that they break confounding associations. For instance, this hand-tying experiment allows us to dissociate the effect of hand use from the potential underlying correlation between left-handedness and evil behavior because, for instance, they're already treated as pariahs and hey why not. In the experiment, people that have and do not have that experience as pariahs can only use their left hand so we can see if it is still associated with evilness.

Final thoughts and cautions

Finally, let me note a few things. First, the Bradford Hill criteria are all about *data or evidence*. That is, when we think about associations we mean associations between a putative cause and an effect across multiple individuals or places or whatever, or perhaps over multiple time points or trials. A single cause and effect story is just an anecdote[19].

[19] "This one time, in band camp, a friend told a joke about squirrels and Kool-aid came out of my nose!" is different than, "Every middle school cafeteria worker I've talked with, and there are now dozens, has a story about a kid laughing so hard milk came out of their nose!" The first is clearly an anecdote, while the second is *approaching* data.

Related to this is that we are talking about observations of things that *have* happened, not our expectations of what is likely to happen. That is called a prediction, and it is not (yet) data or evidence.

Similarly, it is important to remember that we are talking about *causal* relationships, as in a change in x *causes* a change in y. Clouds and lightning are clearly associated, but it wouldn't be right to say that clouds *cause* lightening. Sure they're involved, but it's really the forces at work within clouds that lead to lightening. It can be very helpful to draw out the causal diagram you are considering (e.g., $x \rightarrow z \rightarrow y$) to keep track of what you think *causes* what else (e.g., a change in x

causes a change in z, which causes a change in y), and how it is thought to lead to the response (e.g., changes in x cause changes in y *through its influence* on z, but not directly).

Perhaps most importantly, we are using these "criteria" to see what each line of evidence brings to the table in terms support for or against the causal nature of the relationship. That is, it helps us see that we have a lot of one sort of evidence, but very little of others, or, alternatively, that we might have different stories emerging out of different sorts of evidence. The BHC help us make sense of the evidence we can bring to bear about the causal relationship. It should not be viewed as a checklist or a simple scoring system, but as a tool to help us think through what we do and do not know and why.

Lastly, most of these words have other, perhaps more common usage. We are interested in their meaning about data or evidence, not how you would use them in other settings. For instance, we might think about specific instructions on how to write out a wish-list for your birthday or getting the very specific present you were hoping to get, but neither of those is relevant to thinking about evidence. Instead, specificity applies to the where the putative cause and effect relationship shows up, like only those people that followed these (quite detailed) instructions wound up getting the very present they most fervently hoped to get, whereas those that didn't follow the instructions got bupkis.

So now, go back and re-read those "criteria" with this broader view of what they are and how they are meant to be used. With practice and an open mind these little gems of Bradford Hill will be useful tools.